Lecture 18: November 2

Lecturer: Alessandro Rinaldo *Scribe: YJ Choe*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

18.1 Dudley's integral entropy bound

In this lecture, we introduce an integral bound that gives one of the sharpest bounds on the expected supremum of sub-Gaussian processes. This integral bound can be useful for computing concentration bounds on infinite-dimensional function spaces with known metric entropy.

Theorem 18.1 (Dudley's integral entropy bound)**.** *Let $\{X_\theta : \theta \in \mathbb{T}\}$ be a zero-mean sub-Gaussian process with metric d on the set $\mathbb{T}$. Suppose that*

$$D = \sup_{\theta,\theta' \in \mathbb{T}} d(\theta, \theta') < \infty$$

Then, for any $\delta \in [0, D]$,

$$\mathbb{E}\left[\sup_{\theta,\theta' \in \mathbb{T}} (X_\theta - X_{\theta'})\right] \le 2\,\mathbb{E}\left[\sup_{\substack{\gamma,\gamma' \in \mathbb{T} \\ d(\gamma,\gamma') \le \delta}} (X_\gamma - X_{\gamma'})\right] + 16\mathcal{J}\left(\delta/4, \mathbb{T}\right)$$

where

$$\mathcal{J}(\delta, \mathbb{T}) = \int_\delta^D \sqrt{\log N(\mu, \mathbb{T})}d\mu$$

is the δ-truncated Dudley's entropy integral.

Remark 18.2. (1) Constants in the upper bound can be improved.

(2) The same result holds for $|X_\theta - X_{\theta'}|$ and $|X_\gamma - X_{\gamma'}|$, up to constants.

(3) Typically, we let $\delta \to 0$ to use the following simplified bound:

$$\mathbb{E}\left[\sup_{\theta,\theta' \in \mathbb{T}} (X_\theta - X_{\theta'})\right] \le C \int_0^D \sqrt{\log N(\mu, \mathbb{T})}d\mu$$

for some $C > 0$.

(4) The theorem also gives a bound on $\mathbb{E}\left[\sup_{\theta \in \mathbb{T}} X_\theta\right]$, which is bounded from above by $\mathbb{E}\left[\sup_{\theta,\theta' \in \mathbb{T}} (X_\theta - X_{\theta'})\right]$.

Proof: We start with the 1-step discretization bound:

$$\sup_{\theta,\theta\in\mathbb{T}}(X_\theta - X_{\theta'}) \leq 2 \sup_{\substack{\gamma,\gamma'\in\mathbb{T}\\ d(\gamma,\gamma')\leq\delta}}(X_\gamma - X_{\gamma'}) + 2\max_{i,=1,\ldots,N}|X_{\theta_i} - X_{\theta_1}|$$

where N is the δ-covering number of $\mathbb{T}$ and $\mathbb{U} = \{\theta_1,\ldots,\theta_N\} \subseteq \mathbb{T}$ is a minimal δ-covering of $\mathbb{T}$. Taking expectations give the first term in the upper bound, so we have to bound the second term in expectation.

For each $m = 1, 2, \ldots$, define $\epsilon_m = D \cdot 2^{-m}$ and let $\mathbb{U}_m \subseteq \mathbb{T}$ be a minimal ϵ_m-covering of $\mathbb{U}$ from $\mathbb{T}$. Then, since $\mathbb{U}$ is finite and ϵ_m is decreasing, we can choose L to be the smallest integer such that $|\mathbb{U}_L| = N$. In such case, ϵ_L must be sufficiently small[1], so we can choose $\mathbb{U}_L = \mathbb{U}$.

Note that the choice of L implies that the norm balls $B(\theta_i, \epsilon_L)$ do not intersect for any $i = 1,\ldots,N$, i.e.

$$d(\theta,\theta') > \epsilon_L = D\cdot 2^{-L} \qquad \forall\theta,\theta'\in\mathbb{U}$$

Since L is the *smallest* such integer, we know that there exists some $\theta,\theta'\in\mathbb{U}$ such that $d(\theta,\theta')\leq\epsilon_{L-1} = D\cdot 2^{-(L-1)}$ (otherwise, $L-1$ will be the smallest integer instead). At the same time, we know that $\mathbb{U}$ is a δ-covering of $\mathbb{T}$, so that for such θ,θ',

$$\delta < d(\theta,\theta') \leq D\cdot 2^{-(L-1)}$$

We will use this relationship between δ and L towards the end of the proof.

For each $m = 1,\ldots,L$, define the mapping $\pi_m : \mathbb{U}\to\mathbb{U}_m$ as

$$\pi_m(\theta) = \operatorname*{argmin}_{\beta\in\mathbb{U}_m} d(\theta,\beta)$$

i.e. the best approximation of $\theta\in\mathbb{U}$ from $\mathbb{U}_m$.

Next, for each $\theta\in\mathbb{U}$, let $(\gamma_1,\ldots,\gamma_L)$ be a sequence of points in $\mathbb{T}$ such that $\gamma_L = \theta$ and

$$\gamma_m = \pi_m(\gamma_{m+1})$$

for $m = 1,\ldots,L-1$. We call this sequence a *chain*, as we have the *chaining relation*

$$X_\theta - X_{\gamma_1} = X_{\gamma_L} - X_{\gamma_1} = \sum_{m=2}^{L}\left(X_{\gamma_m} - X_{\gamma_{m-1}}\right)$$

By triangle inequality,

$$|X_\theta - X_{\gamma_1}| \leq \sum_{m=2}^{L}\left|X_{\gamma_m} - X_{\gamma_{m-1}}\right|$$

Given another $\theta'\in\mathbb{U}$, we can construct another chain γ' such that

$$\left|X_{\theta'} - X_{\gamma'_1}\right| \leq \sum_{m=2}^{L}\left|X_{\gamma'_m} - X_{\gamma'_{m-1}}\right|$$

Then, for any $\theta,\theta'\in\mathbb{U}$,

$$\begin{aligned}|X_\theta - X_{\theta'}| &\leq \left|X_{\gamma_1} - X_{\gamma'_1}\right| + |X_\theta - X_{\gamma_1}| + \left|X_{\theta'} - X_{\gamma'_1}\right| \\ &= \left|X_{\gamma_1} - X_{\gamma'_1}\right| + \sum_{m=2}^{L}\left|X_{\gamma_m} - X_{\gamma_{m-1}}\right| + \sum_{m=2}^{L}\left|X_{\gamma'_m} - X_{\gamma'_{m-1}}\right|\end{aligned}$$

[1] ϵ_L must be small enough such that $d(\theta_i,\theta_{i'}) > \epsilon_L$ for all $i\neq i'$.

and each of the two alternating sums is bounded by

$$\sum_{m=2}^{L} \max_{\beta \in \mathbb{U}_m} \left| X_\beta - X_{\pi_{m-1}(\beta)} \right|$$

Then, taking maximum over all $\theta, \theta' \in \mathbb{U}$ and expectations, we get

$$\mathbb{E}\left[\max_{\theta,\theta' \in \mathbb{U}} \left| X_\theta - X_{\tilde{\theta}} \right|\right] \leq \mathbb{E}\left[\max_{\gamma,\gamma' \in \mathbb{U}_1} \left| X_\gamma - X_{\tilde{\gamma}} \right|\right] + 2\sum_{m=2}^{L} \mathbb{E}\left[\max_{\beta \in \mathbb{U}_m} \left| X_\beta - X_{\pi_{m-1}(\beta)} \right|\right]$$

To bound the first term, notice that

$$X_\gamma - X_{\gamma'} \in SG\left(d^2(\gamma, \gamma')\right)$$

Since $d(\gamma, \tilde{\gamma}) \leq D = \sup_{\theta,\theta' \in \mathbb{T}} d(\theta, \theta') < \infty$ by assumption, we get

$$X_\gamma - X_{\gamma'} \in SG\left(D^2\right)$$

Then, by the metric entropy bound for sub-Gaussian random variables, we get

$$\mathbb{E}\left[\max_{\gamma,\gamma' \in \mathbb{U}_1} \left| X_\gamma - X_{\gamma'} \right|\right] \leq 2D\sqrt{\log N(D/2, \mathbb{T})}$$

where we recall that $\mathbb{U}_1$ is a minimal ϵ_1-covering of $\mathbb{U} \subseteq \mathbb{T}$ and $\epsilon_1 = D \cdot 2^{-1} = D/2$, so that $|\mathbb{U}_1| \leq N(D/2, \mathbb{T})$ by definition.

To bound the second term, for $m = 2, \ldots, L$, we can give an analogous metric entropy bound. First, we have that

$$\max_{\beta \in \mathbb{U}_m} d(\beta, \pi_{m-1}(\beta)) \leq D \cdot 2^{-(m-1)}$$

since π_{m-1} is the best approximation from $\mathbb{U}_{m-1}$. Also,

$$|\mathbb{U}_m| \leq N(D \cdot 2^{-m}, \mathbb{T})$$

because $\mathbb{U}_m$ is a minimal ϵ_m-covering of $\mathbb{U}$ where $\epsilon_m = D \cdot 2^{-m}$. Thus, we get

$$\mathbb{E}\left[\max_{\beta \in \mathbb{U}_m} \left| X_\beta - X_{\pi_{m-1}(\beta)} \right|\right] \leq 2 \cdot D \cdot 2^{-(m-1)} \sqrt{\log N(D \cdot 2^{-m}, \mathbb{T})}$$

for $m = 2, \ldots, L$.

Combining these bounds on the two terms, we have

$$\mathbb{E}\left[\max_{\theta,\theta' \in \mathbb{U}} (X_\theta - X_{\theta'})\right] \leq 4 \sum_{m=1}^{L} \left[D \cdot 2^{-(m-1)} \sqrt{\log N(D \cdot 2^{-m}, \mathbb{T})}\right]$$

This already is a good bound, but we can also bound it with an integral, using the fact that $\mu \mapsto \sqrt{\log N(\mu, \mathbb{T})}$ is a non-increasing function on $[0, D]$. Because the function is non-increasing, the summation is a lower

Riemann approximation of the integral of $\mu \mapsto \sqrt{\log N(\mu, \mathbb{T})}$. That is,

$$\begin{aligned}
\mathbb{E}\left[\max_{\theta,\theta' \in \mathbb{U}} (X_\theta - X_{\theta'})\right] &\leq 4 \sum_{m=1}^{L} \left[D \cdot 2^{-(m-1)} \sqrt{\log N(D \cdot 2^{-m}, \mathbb{T})}\right] \\
&\leq 4 \sum_{m=1}^{L} \left[2 \int_{D \cdot 2^{-(m+1)}}^{D \cdot 2^{-m}} \sqrt{\log N(\mu, \mathbb{T})} d\mu\right] \\
&\leq 8 \int_{D/2^{L+1}}^{D/2} \sqrt{\log N(\mu, \mathbb{T})} d\mu \\
&\leq 8 \int_{\delta/4}^{D} \sqrt{\log N(\mu, \mathbb{T})} d\mu \\
&= 8 \mathcal{J}(\delta/4, \mathbb{T})
\end{aligned}$$

where for the last inequality we use the fact that $\delta/4 \leq D/2^{L+1}$, which follows from what we derived earlier that $\delta \leq D \cdot 2^{-(L-1)}$. Plugging this result into the 1-step discretization bound, we get

$$\begin{aligned}
\mathbb{E}\left[\sup_{\theta,\theta' \in \mathbb{T}} (X_\theta - X_{\theta'})\right] &\leq 2\, \mathbb{E}\left[\sup_{\substack{\gamma,\gamma' \in \mathbb{T} \\ d(\gamma,\gamma') \leq \delta}} (X_\gamma - X_{\gamma'})\right] + 2\, \mathbb{E}\left[\max_{i,=1,\dots,N} |X_{\theta_i} - X_{\theta_1}|\right] \\
&\leq 2\, \mathbb{E}\left[\sup_{\substack{\gamma,\gamma' \in \mathbb{T} \\ d(\gamma,\gamma') \leq \delta}} (X_\gamma - X_{\gamma'})\right] + 2\, \mathbb{E}\left[\max_{\theta,\theta' \in \mathbb{U}} \left(X_\theta - X_{\tilde{\theta}}\right)\right] \\
&\leq 2\, \mathbb{E}\left[\sup_{\substack{\gamma,\gamma' \in \mathbb{T} \\ d(\gamma,\gamma') \leq \delta}} (X_\gamma - X_{\gamma'})\right] + 16 \mathcal{J}(\delta/4, \mathbb{T})
\end{aligned}$$

■

Example 18.3 (Uniform bounds on VC classes)**.** Let $\mathcal{F}$ be a function class on $\mathcal{X}$ with VC-dimension ν. (For example, $\mathcal{F} = \{(-\infty, x] : x \in \mathbb{R}\}$ with $VC(\mathcal{F}) = 1$.) We saw earlier in the course that bounding

$$\|P_n - P\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}[f] \right|$$

can be reduced via symmetrization to bounding the empirical Rademacher complexity

$$\mathcal{R}_n(\mathcal{F}, x^n) = \mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^{n} \epsilon_i f(x_i) \right|\right]$$

for $x^n = (x_1, \dots, x_n)$ and Rademacher random variables $\epsilon_1, \dots, \epsilon_n \in \{-1, +1\}$.

Fix any x^n, and define

$$Z_f = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i f(x_i)$$

for each $f \in \mathbb{F}$. Then, for any $f, g \in \mathcal{F}$,

$$Z_f - Z_g \in SG\left(\|f - g\|_n^2\right)$$

so that $\{Z_f\}_{f\in\mathcal{F}}$ is a sub-Gaussian process with the metric $d(f,g) = \|f-g\|_n = \sqrt{\frac{1}{n}\sum_{i=1}^n (f(x_i)-g(x_i))^2}$. Using Dudley's entropy integral bound, we immediately get

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\left|\sum_{i=1}^n \epsilon_i f(x_i)\right|\right] \le \frac{C}{\sqrt{n}}\int_0^2 \sqrt{\log N(\mu,\mathcal{F},\|\cdot\|_n)}d\mu$$

To bound the metric entropy on the right-hand side using the VC dimension, we use the following theorem.

Theorem 18.4 (2.6.7 in [VW06]). *Let $\mathcal{F}$ be a function class on $\mathcal{X}$ with VC-dimension ν. Assume that $\mathcal{F}$ is uniformly bounded by $b>0$. Then, for any probability distribution Q on $\mathcal{X}$ and for any $p\ge 1$,*

$$N(\delta,\mathcal{F},\|\cdot\|_{L_p(Q)}) \le C_0\cdot(\nu+1)(16e)^{\nu+1}\left(\frac{b}{\delta}\right)^{p\nu}$$

for some universal constant $C_0>0$, where for any $f,g\in\mathcal{F}$,

$$\|f-g\|_{L_p(Q)} = \left(\int |f-g|^p\, dQ\right)^{1/p}$$

Then, Dudley's entropy integral bound becomes

$$\begin{aligned}\mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\left|\sum_{i=1}^n \epsilon_i f(x_i)\right|\right] &\le \frac{C}{\sqrt{n}}\int_0^2 \sqrt{\log N(\mu,\mathcal{F},\|\cdot\|_n)}d\mu \\ &\le C'\cdot\sqrt{\frac{\nu}{n}}\int_0^{2b}\sqrt{\log(b/\delta)}d\mu \\ &\lesssim \sqrt{\frac{\nu}{n}}\end{aligned}$$

This is a sharper result than our previous VC result, which gives the rate $\sqrt{\frac{\nu\log n}{n}}$. In general, results using Dudley's bound can be more powerful and do not require VC concentration bounds.

References

[VW06] A. W. van der Vaart, J. A. Wellner, "Weak Convergence and Empirical Processes," *Springer Series in Statistics*, 2006.